Acquisition temps-réel de données articulatoires par IRM : application à la synthèse par copie

Benjamin Elie, Yves Laprie, Pierre-André Vuissoz

CFA/VISHNO 2016

Acquisition temps-réel de données articulatoires par IRM : application à la synthèse par copie

B. Elie[a], Y. Laprie[a] et P.-A. Vuissoz[b]
[a]Loria, Campus Scientifique, 615 Rue du Jardin Botanique, 54506 Vandœuvre-Lès-Nancy, France
[b]Imagerie Adaptative Diagnostique et Interventionnelle, CHU de Nancy Brabois - Tour Drouet - Rue du Morvan, 54511 Vandoeuvre-Lès-Nancy, France
benjamin.elie@inria.fr

2009

L'étude de la production de la parole nécessite de connaitre précisément l'évolution temporelle de la géométrie du conduit vocal. Récemment, l'imagerie par résonance magnétique (IRM) a été couramment utilisée car elle possède les avantages d'offrir une image des tissus internes avec un fort contraste, de sélectionner des coupes précises, tout en étant inoffensive pour les sujets. Toutefois, les contraintes sur le temps d'acquisition peuvent être un obstacle majeur pour l'acquisition en temps-réel des mouvements articulatoires rapides du fait d'une cadence d'acquisition trop faible. Cette étude présente une méthode permettant d'accélérer la cadence d'acquisition, jusqu'à obtenir 36 images par seconde, grâce à l'acquisition compressée (ou Compressed Sensing). En effet, en utilisant l'a priori d'une forte parcimonie de la transformée de Fourier temporelle de la séquence d'image, et en choisissant un parcours adapté dans l'espace d'acquisition, il est alors possible de réduire considérablement le nombre d'observations dans l'espace d'acquisition, tout en garantissant une reconstruction satisfaisante de la séquence d'image pour leur post-traitement. Celui-ci est effectué à l'aide d'une reconnaissance des contours des différents articulateurs, ainsi que de l'utilisation d'un modèle articulatoire adapté, ce qui permet alors d'extraire l'évolution temporelle des paramètres du modèle articulatoire, ainsi que les fonctions d'aire correspondantes. Nous montrons également que ces données peuvent alors servir de paramètres d'entrée de synthétiseurs acoustiques dans le but d'analyser le lien entre les mouvements articulatoires du locuteur et les indices acoustiques de la parole.

1 Introduction

Les études sur la production de la parole requièrent la connaissance simultanée des enregistrements audio de la parole et les données articulatoires dans le but de relier les indices acoustiques contenus dans la parole naturelle à leurs origines articulatoires. Du fait de nombreuses contraintes (risques sanitaires, accès aux tissus internes…), l'acquisition des données articulatoires est difficile et pose des problèmes majeurs.

La production de la parole implique deux niveaux de coordination. Les articulateurs de la parole doivent se mouvoir en fonction de la séquence de phonèmes à articuler (coarticulation). De plus, la réalisation de constrictions supra-glottiques et le couplage avec la cavité nasale doit être coordonnée avec la configuration de la glotte afin que les propriétés aérodynamiques du conduit vocal soient compatibles avec la nature de la source d'excitation. Ces deux mécanismes de coordination s'opèrent avec des échelles de temps différentes. Au niveau segmental, à savoir celui des sons de la parole, l'ordre de grandeur de la coordination entre les cavités supra-glottiques et les plis vocaux est d'une milliseconde, alors que celui de la coordination entre les articulateurs supra-glottiques est d'une centaine de millisecondes. Cela implique que l'étude de la coarticulation, qui est notre principal objectif, requière une acquisition de données articulatoires à une cadence supérieure à 10 Hz, voire au moins 30 Hz pour obtenir des données articulatoires pouvant être exploitées. En effet, cela donne approximativement au moins une à deux images pour chaque phone.

Récemment, l'*Imagerie par Résonance Magnétique* (IRM) a suscité un vive intérêt pour l'acquisition de données articulatoires de par les nombreux avantages qu'elles présentent [1, 2, 3]. Contrairement aux rayons X, elles sont inoffensives pour le sujet, elles permettent une visualisation 3D des tissus internes, ou tout simplement une seule couche, sans subir de perturbations dues aux couches adjacentes. Cependant, elles souffrent d'un inconvénient majeur dû à la lenteur d'acquisition. En effet, acquérir l'intégralité de l'espace k (ou espace de Fourier) correspondant à une seule coupe peut prendre plusieurs centaines de millisecondes, et par conséquent, contraindre fortement la résolution temporelle à être au-dessus de cette valeur, ce qui n'est pas suffisant pour atteindre les objectifs de cadence d'acquisition cités précédemment.

Historiquement, l'IRM appliquée à la parole a utilisé un schéma d'échantillonnage de l'espace k en forme de spirale afin d'accélérer la cadence d'acquisition [1]. Ce schéma d'échantillonnage offre une image de bonne qualité étant donné la grande cadence d'acquisition. Cependant, cela peut générer de forts artéfacts non désirés, telles que des déformations non réalistes de la pointe de la langue lors de la rétrofléxion de celle-ci, ou une élongation exagérée des lèvres. Cela perturbe alors considérablement l'estimation des contours des articulateurs, et par conséquent, l'acquisition de données articulatoires. Afin d'éviter de tels artéfacts, la méthode proposée dans cet article utilise une modification du schéma d'échantillonnage de type cartésien, à savoir une acquisition ligne par ligne de l'espace k. Le choix du schéma cartésien est également motivé par sa faculté à être utilisé avec l'acquisition compressée (CS, pour *Compressed Sensing*) [4] simultanément avec une reconstruction homodyne [5]. Ces paradigmes mathématiques permettent la reconstruction d'images à partir d'observations partielles de l'espace d'acquisition. Cet article présente une méthode pour simultanément intégrer plusieurs techniques d'accélération utilisées en IRM, et l'appliquer à l'acquisition de données articulatoires. Le paragraphe 2.1 détaille les aspects théoriques liés à l'acquisition compressée et aux reconstructions homodynes, ainsi que le choix du schéma d'échantillonnage. Des simulations et des expériences sur volontaires sains sont présentées au paragraphe 3, et les résultats au paragraphe 4. Les données articulatoires sont collectées à partir des films IRM obtenus suivant la méthode détaillée au paragraphe 5.

2 Théorie

2.1 Acquisition compressée

L'acquisition compressée (CS) est une méthode mathématique permettant à un signal échantillonné à une cadence qui ne respecte pas les conditions de Shannon [4] d'être reconstruit par la résolution exacte du problème inverse consistant à retrouver les données manquantes. La fiabilité de reconstruction de CS repose sur l'hypothèse que le signal $\mathbf{x} \in \mathbb{C}^n$ à reconstruire est K-parcimonieux dans une certaine base, c'est-à-dire qu'il existe une transformée parcimonieuse $\mathbf{\Psi}$ telle que $\mathbf{\Psi x} \in \mathbb{C}^n$ ne contient que $K < n$ éléments non-nuls. Dans ce cas, en présence de bruit, seulement $m \geq K$ observations de $\mathbf{x}$ suffisent pour trouver la

solution du problème convexe

$$\mathbf{x} = \underset{\hat{\mathbf{x}}}{\operatorname{argmin}} \|\mathbf{\Psi}\hat{\mathbf{x}}\|_1 \quad \text{s.t.} \quad \|\mathbf{\Phi}\hat{\mathbf{x}} - \mathbf{b}\|_2 \leq \epsilon, \tag{1}$$

où $\mathbf{\Phi} \in \mathbb{R}^{m \times n}$ est la matrice d'encodage qui ne contient que des 0 et des 1, telle que $\mathbf{\Phi x} = \mathbf{b}$, et $\mathbf{b} \in \mathbb{C}^m$ est le signal observé, à savoir la version sous-échantillonnée de $\mathbf{x}$, et ϵ est une valeur de tolérance au bruit de mesure. La norme ℓ_1 $\|\mathbf{x}\|_1$ du vecteur $\mathbf{x}$ est définie en tant que somme des modules des éléments complexes de $\mathbf{x}$. Une autre condition importante pour une reconstruction exacte est l'incohérence de la matrice d'encodage $\mathbf{\Phi}$ avec la transformée parcimonieuse $\mathbf{\Psi}$.

Lorsque l'on applique cette technique à des données IRM, le signal désiré est généralement l'image des tissus internes, à savoir la transformée de Fourier spatiale inverse de $\mathbf{x}$, notée $\mathcal{F}_{sp}^{-1}\mathbf{x}$. Dans notre cas, la transformée parcimonieuse $\mathbf{\Psi}$ a été choisie comme la transformée de Fourier temporelle des images, à savoir

$$\mathbf{\Psi} = \mathcal{F}_t \mathcal{F}_{sp}^{-1}\mathbf{x},$$

où $\mathcal{F}_t$ est la transformée de Fourier temporelle. Cela se justifie par le fait que la grande partie de l'image ne bouge pas, et que seulement une petite portion de ces images représente des mouvements relativement lents. En conséquence, la plupart des pixels possède une transformée de Fourier temporelle très parcimonieuse. En comparaison avec des transformées de type ondelettes, que nous avons essayées, la transformée de Fourier temporelle des images permet d'obtenir de meilleurs contrastes aux frontières du conduit vocal et des articulateurs. En pratique, Eq. 1 est alors

$$\rho = \underset{\hat{\rho}}{\operatorname{argmin}} \|\mathcal{F}_t\hat{\rho}\|_1 \quad \text{s.t.} \quad \|\mathbf{\Phi}\mathcal{F}_{sp}\hat{\rho} - \mathbf{b}\|_2 \leq \epsilon, \tag{2}$$

où ρ est l'ensemble d'images à reconstruire.

2.2 Acquisition compressée jointe

En pratique, l'IRM utilise également des antennes multicanaux. Elles sont communément utilisées pour accélérer la cadence d'acquisition à l'aide des techniques d'imagerie parallèle, telles que SENSE (*SENSitivity Encoding*) [6], ou GRAPPA (*Generalized Autocalibrating Partially Paralleled Acquisitions*) [7]. Ces techniques se basent sur une acquisition partielle de l'espace k à l'aide d'un sous-échantillonnage régulier, puis d'une reconstruction de l'image repliée à l'aide des informations spatiales inclues dans les différentes antennes. Pour notre cas, cela n'est pas forcément intéressant car le repliement des images diminue la parcimonie de notre signal. En revanche, la disponibilité de plusieurs canaux peut être exploité à l'aide de l'acquisition compressée jointe (DCS pour *Distributed Compressed Sensing*). Cela s'effectue en exploitant l'idée que les signaux reçus par les différentes antennes sont fortement corrélés. En effet, leur parcimonie dans la base parcimonieuse choisie est similaire. Par conséquent, il est possible d'introduire l'a priori de parcimonie jointe pour régulariser le problème inverse, en minimisant simultanément la norme ℓ_1 de la représentation parcimonieuse du signal de chaque antenne, et le nombre de lignes non-nulles de la matrice signal (la matrice dont les colonnes contiennent le signaux reçus dans chaque antenne). Cette contrainte impose donc que la position des coefficients non-nuls de la représentation parcimonieuse soit identique d'une antenne à l'autre. Le problème DCS [8] s'écrit alors

$$\mathbf{P} = \underset{\hat{\mathbf{P}}}{\operatorname{argmin}} \|\mathcal{F}_i\hat{\mathbf{P}}\|_{1,2} \quad \text{s.t.} \quad \|\mathbf{\Phi}\mathcal{F}_{sp}\hat{\mathbf{P}} - \mathbf{B}\|_{2,2} \leq \epsilon, \tag{3}$$

où $\mathbf{P} \in \mathbb{C}^{n \times l}$ est la matrice dont les colonnes contiennent la représentation en vecteur des images reconstruites dans chacune des l antennes, et $\mathbf{B} \in \mathbb{C}^{m \times l}$ est la matrice d'observation. La norme mixte $\ell_{1,2}$ de la matrice $\mathbf{P}$ est définie comme la norme ℓ_1 de la norme ℓ_2 de chaque ligne de $\mathbf{P}$, d'où

$$\|\mathbf{P}\|_{1,2} = \sum_{i=1}^{n} \|\mathbf{P}_i\|_2, \tag{4}$$

où $\mathbf{P}_i$ est la $i^{\text{ème}}$ ligne de $\mathbf{P}$.

2.3 Reconstruction homodyne

Les techniques de reconstruction homodyne sont basées sur l'hypothèse que l'espace k possède la propriété de symétrie hermitienne du fait que les images à reconstruire sont à valeurs réelles. Par conséquent, la connaissance de seulement la moitié de l'espace de Fourier suffit. En pratique, cette propriété n'est plus valide car des erreurs de phase viennent modifier la propriété de symétrie hermitienne de l'espace k. Il est cependant possible d'effectuer des corrections de phase à partir d'un échantillonnage partiel de l'espace k lorsque la portion de l'espace observé est supérieure à 1/2.

La reconstruction homodyne [9] utilise une correction de phase définie par

$$p^*(x, y) = e^{-j\angle\rho_{lr}(x,y)}, \tag{5}$$

où ρ_{lr} est une version basse-résolution de l'image ρ. C'est la transformée de Fourier spatiale inverse du signal observé fenêtré par une fonction porte centrée autour des très basses fréquences. La multiplication de l'image ρ_i obtenue par la transformée de Fourier spatiale inverse de l'espace k observé (dont les données partiellement acquises sont complétés par des zéros) par $p^*(x, y)$ permet donc d'obtenir une image corrigée en terme de phase. Parmi les techniques de reconstruction homodyne, POCS (*Projection Onto Convex Sets*) [5] s'est révélée performante pour des fractions petites d'observations de l'espace k. Elle consiste en une application itérative de la correction de phase jusqu'à ce que les modifications d'une itération à l'autre deviennent négligeable.

En application avec CS, du fait du grand nombre d'informations manquantes, l'image basse-résolution ρ_{lr} est calculée à partir de la moyenne temporelle de l'espace $k - t$ acquis.

2.4 Choix du schéma d'échantillonnage

Les études récentes à propos de l'acquisition de mouvements articulatoires par IRM se sont focalisées sur le schéma d'échantillonnage en forme de spirale [2, 3]. Quoique très efficace pour accélérer le processus d'acquisition, il n'en reste pas moins sujet à des artéfacts qui peuvent gêner le dépouillement des données. De par sa simplicité d'utilisation et d'implémentation sur des machines IRM temps-réel, un schéma d'échantillonnage pseudo-aléatoire de type cartésien a été préféré pour cette étude. Il a également été prouvé qu'un échantillonnage cartésien à densité variable,

où les lignes centrales (basses fréquences) sont privilégiées, amène de meilleures reconstructions qu'un échantillonnage aléatoire uniforme [10].

Au regard de ces considérations, le schéma d'échantillonnage est défini comme suit. Premièrement, un nombre de lignes à encoder par trame temporelle, noté n_{lpf}, est fixé à une valeur choisie. Cette valeur est un compromis entre une bonne résolution temporelle (faible n_{lpf}) et une meilleure qualité d'image (grand n_{lpf}). Un nombre de lignes basses fréquences, noté n_{cl} avec $n_{cl} \leq n_{lpf}$, sont constamment acquises à chaque trame temporelle. Finalement, les $n_{lpf} - n_{cl}$ lignes restantes à acquérir à la trame temporelle t sont aléatoirement choisies en respectant la fonction de probabilité suivante

$$p(k_y, t) = \left| \frac{1}{\left[1 - (k_y - n_y/2)\right]^{r(t)}} \right|, \qquad (6)$$

où k_y est le numéro de la ligne de l'espace k, $n_y/2$ est la ligne centrale (fréquence nulle), et $r(t) \in [0, 0.5]$ est un nombre choisi aléatoirement à chaque trame temporelle en respectant une distribution aléatoire uniforme. Cette fonction de probabilité donne alors une densité d'échantillonnage qui décroît à mesure que l'on s'éloigne des basses fréquences. Un exemple de trajectoire d'échantillonnage est donné en figure 1.

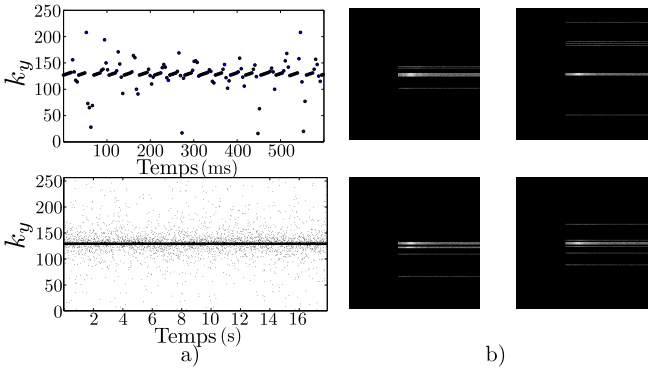

FIGURE 1 – Trajectoire d'échantillonnage utilisé pour l'étude. Colonne de gauche : schéma d'échantillonnage séquentiel des lignes encodées en fonction du temps : (haut) détail ligne par ligne, (bas) schéma global, trame par trame. Droite : 4 premiers espaces k acquis. À noter que chaque ligne n'est pas entièrement encodée : les données manquantes sont reconstruites en utilisant POCS [5].

2.5 Méthode de reconstruction proposée

Pour résumer, le protocole expérimental pour acquérir des données articulatoires avec une haute résolution spatio-temporelle est défini comme suit

1. choix de la coupe à observer (généralement la coupe médio-sagittale),
2. choix des paramètres d'acquisition (n_y, n_x, n_{lpf}, et n_{cl}) en fonction de la résolution spatiotemporelle désirée,
3. calcul du schéma d'échantillonnage en utilisant l'Eq. (6),
4. reconstruction des données manquantes des lignes partiellement acquises en utilisant POCS [5]
5. reconstruction DCS en utilisant l'Eq. (4) et l'algorithme d'optimisation ℓ_1 SPGL1 [11, 12],
6. processus de débruitage pour améliorer la qualité des images reconstruites en utilisant la méthode détaillée dans [13],
7. recombinaison des données des différentes antennes pour les convertir en une seule antenne combinée [14].

3 Matériel et méthode pour l'acquisition de données articulatoires

3.1 Simulations

Des simulations avec fantômes numériques sont effectuées pour évaluer la qualité de reconstruction de la méthode présentée ici en fonction du nombre de lignes par trame n_{lpf}, qui est lié à la cadence d'acquisition obtenue. Les fantômes numériques sont créés à partir d'images de coupes médio-sagittales obtenues par IRM avec une technique indépendante [15]. Les séquences d'images sont alors interpolées sur un vecteur temps aléatoire, où chaque pas de temps correspond au temps d'acquisition d'une seule ligne, dans le but de simuler des accélération et ralentissement des gestes articulatoires. L'espace $k - t$ correspondant est alors sous-échantillonné au regard des différents schémas d'échantillonnage évalués. La reconstruction DCS est alors appliquée aux données simulées. Les images sont de tailles 256×256 pixels, correspondant à 1.016×1.016 mm^2. Chaque séquence contient 128 images simulées sur 16 antennes. Le pas temporel est fixé à celui du temps de répétition utilisé en pratique, à savoir 3.5 ms, et n_{cl} est fixé à 5.

La figure 2 représente les valeurs NRMSE *Normalized Root Mean Square Error*) des images reconstruites comparées aux fantômes numériques en fonction du nombre de lignes par trame n_{lpf}. Pour calculer les valeurs NRMSE, seuls les pixels inclus dans la région d'intérêt (autour du conduit vocal). Tel que l'on pouvait s'y attendre, la qualité de reconstruction est d'autant meilleur que n_{lpf} est grand. Cependant, pour $n_{lpf} > 10$, l'augmentation de la qualité de reconstruction n'est plus très significative. Par conséquent, fixer n_{lpf} à 10 en pratique est un bon compromis entre une haute résolution temporelle et une bonne qualité d'image.

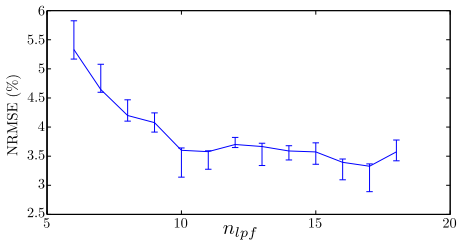

FIGURE 2 – Valeur médiane NRMSE des images reconstruites en fonction de n_{lpf}. Les barres d'erreurs indiquent les écarts absolus médians par valeurs supérieures et inférieures à la médiane. 5 réalisations sont effectuées pour chaque valeur de n_{lpf}.

3.2 Expériences d'IRM

Les expérimentations IRM ont été effectuées sur un système 3T signa HDxt MR (General Electric Healthcare, Milwaukee, WI). Les données IRM articulatoires ont été acquises sur 2 sujets sains avec un consentement écrit et une approbation du comité local d'éthique. Les données ont été collectées sur une antenne neurovasculaire 16 canaux. Le protocole consistait en une coupe médio-sagittale du conduit vocal acquises avec une séquence modifiée de type écho de gradient Spoiled Fast Gradient Echo (Fast SPGR, TR 3.5ms, TE 1.1ms, bande passante 83.33 kHz, angle de rotation 30 degrés, matrice 256×256, 512 trames temporelles).

4 Résultats IRM

La figure 3 trace l'évolution de l'ouverture aux lèvres en fonction du temps durant la prononciation de "J'ai pigé la phrase" (/ʒe.pi.ʒe.la.fʁɑz/), ainsi que le mouvement du dos de la langue. Les boites sur les tracés de droite indiquent les moments de la production du mot (/fʁɑz/).

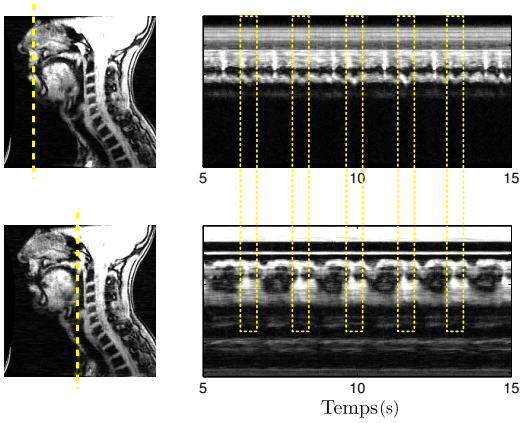

FIGURE 3 – Tracés en bande de répétitions de la phrase "J'ai pigé la phrase" (/ʒe.pi.ʒe.la.fʁɑz/). Haut : tracé en bande au niveau des lèvres. Bas : tracé en bande de la constriction uvulaire. Les positions des coupes sont indiquées par la ligne verticale sur les coupes médio-sagittales à gauche. Les boites sur les tracés de droite indiquent les moments de la production du mot "phrase" (/fʁɑz/). La résolution temporelle est de 35 ms. La résolution spatiale est de 1.016×1.016 mm^2, et l'épaisseur de coupe est de 5 mm.

La résolution temporelle et la qualité d'image sont suffisantes pour visualiser les mouvements des articulateurs. Par exemple, l'ouverture des lèvres augmente pour prononcer /fʁɑz/ : cela va de la constriction étroite pour la fricative labio-dentale /f/ à une ouverture large pour la voyelle ouverte /ɑ/, puis décroit de nouveau pour la fricative alvéolaire voisée /z/. De manière coordonnée, comme on peut le constater sur le tracé en bande du bas de la figure 3, le dos de la langue se relève pour créer une constriction uvulaire caractéristique de la fricative uvulaire voisée /ʁ/, correspondant au "r" français.

La figure 4 montre les tracés en bande de répétition du logatome /ara/, contenant le trille uvulaire /ɾ/ ("r" roulé). Le film est reconstruit à une cadence de 48 images par seconde, et la résolution spatiale est de 1.016×1.016 mm^2. Les moments de production de la consonne roulée /ɾ/ sont indiquées par les boites sur les tracés en bande, à droite de la figure 4.

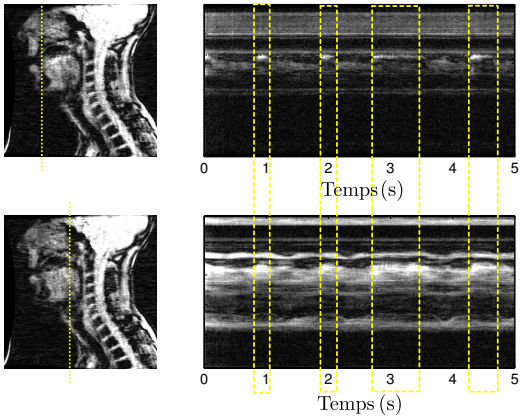

FIGURE 4 – Tracés en bande de répétitions du logatome /ara/. Haut : tracé en bande au niveau alvéolaire. Bas : tracé en bande de la constriction uvulaire. Les positions des coupes sont indiquées par la ligne verticale sur les coupes médio-sagittales à gauche. Les boites sur les tracés de droite indiquent les moments de la production du trille uvulaire /ɾ/. La résolution temporelle est de 20.8 ms. La résolution spatiale est de 1.016×1.016 mm^2, et l'épaisseur de coupe est de 5 mm.

La haute résolution spatiotemporelle permet aux oscillations de la langue d'être observées. Il est aussi intéressant de constater la coarticulation de la pointe de la langue avec son dos : à chaque production du trille /ɾ/, les deux parties de la langue créent une constriction. La constriction uvulaire effectuée avec le dos de la langue permet probablement à la pression intra-orale en amont de la constriction alvéolaire de s'élever, afin que les conditions aérodynamiques au voisinage de la pointe de la langue permettent à celle-ci d'auto-osciller.

5 Exploitation des données articulatoires et synthèse acoustique

5.1 Extraction des données articulatoires

Même si l'acoustique du conduit vocal est déterminé par sa géométrie et les configurations glottiques, détourer les contours de la glotte aux lèvres en un bloc ne suffit pas pour modéliser la coarticulation. En effet, les articulateurs ne bougent d'un bloc d'une position à l'autre. Ils sont organisés en groupes, chacun de ces groupes réalisant un geste particulier. La production de la parole implique des gestes se recouvrant qui doivent être modélisés séparément. Il est donc important de détourer chaque articulateur indépendamment, i.e. la mandibule, la langue, les lèvres, l'épiglotte, le larynx et le vélum.

Nous avons développé un logiciel spécifique, appelé Xarticulators, pour détourer les contours des articulateurs et nous avons exploré deux stratégies pour cela. La première consiste à détourer le contour de la langue à la main. Cela est faisable pour un faible nombre d'images dans le but d'obtenir une bonne qualité de contours et vérifier que les simulations acoustiques reproduisent bien le signal attendu (*cf.* Paragraphe 5.2). Pour une quantité de données plus importante, il est possible de se servir du fait que, contrairement aux rayons X, l'IRM produit des contours qui ne se recouvrent pas. Cette caractéristique est très intéressante en terme de

suivi de contour et nous l'avons exploité en adaptant l'algorithme de suivi semi-automatique développé par Fontecave et Berthommier [16]. Le principe est de détourer les contours à des images-clés, choisies aléatoirement parmi la séquence d'images, et ensuite d'indexer les images où les contours doivent être suivis en fonction des images clés, en utilisant une Transformée en Cosinus Discrète (TCD). L'image utilisée pour indexer les images est limitée à la région où le contour à suivre est localisé. Le contour final est obtenu en moyennant les contours des images les plus proches. Ce suivi semi-automatique requiert la supervision de l'utilisateur qui peut ajouter des images-clés pour corriger les éventuelles erreurs correspondant à une image trop éloignée de l'image-clé existante.

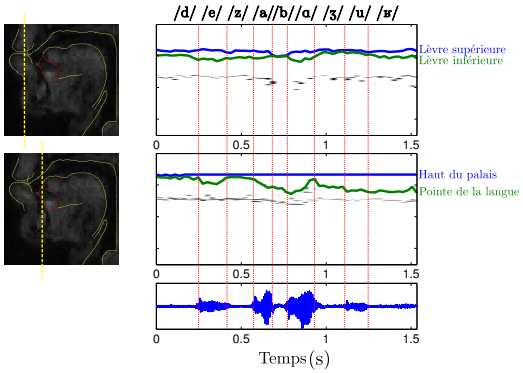

FIGURE 5 – Tracé en bande du contour des articulateurs extrait de la prononciation de la phrase "Des abat-jours" (/de.za.bɑ.ʒuʁ/). Haut : ouverture des lèvres. Centre : ouverture de la région alvéolaire. Bas : signal audio. La segmentation phonétique est indiquée par des lignes traitillées verticales. Colonne de gauche : vue médio-sagittales du conduit vocal avec le tracé des contours des articulateurs et la position correspondant aux tracés en bande (ligne pleine verticale).

La figure 5 montre un exemple de contours d'articulateurs suivis par la méthode présentée. Le suivi correspond à la prononciation de la phrase "Des abat-jours" (/de.za.bɑ.ʒuʁ/), et montre l'ouverture aux lèvres (tracé du haut) et l'ouverture dans la région alvéolaire (tracé central) en fonction du temps. Le signal audio, post-traité pour supprimer le bruit de l'IRM à l'aide d'une méthode de séparation de source [17], est représenté en bas de la figure 5 avec également les indications sur la segmentation phonétique. L'évolution temporelle des contours des articulateurs est en accord avec la segmentation phonétique : une constriction alvéolaire se forme lors de la prononciation de la fricative alvéolaire /z/, les lèvres sont en contact lors de l'occlusive bilabiale /b/, puis se relâchent soudainement pour produire la voyelle ouverte /ɑ/ suivante. Enfin, les lèvres se rapprochent fortement pour prononcer le diphone /ʒu/.

5.2 Synthèse acoustique

Afin de confirmer la pertinence des données articulatoires ainsi extraites à partir des IRM, l'évolution de la géométrie du conduit vocal est donnée en entrée d'un synthétiseur vocal de parole continue [18]. Le synthétiseur consiste en la simulation numérique de la propagation d'une onde plane dans un conduit vocal géométriquement réaliste approximé par sa fonction d'aire, à l'aide du paradigme de la formulation monomatricielle de Mokhtari *et coll.* [19]. A cela s'ajoute un modèle auto-oscillant des plis vocaux de type 2×2 masses [20] avec possibilité de fermeture partielle [18].

Afin de passer d'une représentation 2D du conduit vocal à la fonction d'aire, un algorithme de correction de fonction d'aire à partir des trajectoires formantiques est utilisé [21]. Les fonctions d'aires initiales sont alors celles calculées à l'aide des paramètres α β utilisés dans [22] et des distances médio-sagittales extraites des contours. Afin de préserver une certaine cohérence entre les distances médio-sagittales extraites des contours et les fonctions d'aire corrigées, les termes de pénalisation minimisant la différence entre les fonctions d'aire corrigées et initiales sont fixées à 90 % des termes de pénalisation.

Dans un deuxième temps, les paramètres de la source glottique sont calculées à partir de la fréquence fondamentale et la segmentation phonétiques déduites à partir du signal acoustique acquis simultanément avec les IRM. Cela consiste à, premièrement, modifier le rapport raideur/masse du modèle mécanique des plis vocaux au cours du temps pour que ceux-ci oscillent à la fréquence fondamentale désirée, et deuxièmement, à régler l'abduction partielle des plis vocaux lors de la production des fricatives voisées /z/, /ʒ/, et /ʁ/, pour que celle-ci soit suffisamment élevée afin de garantir à la fois la production de bruit de friction et la source de voisement.

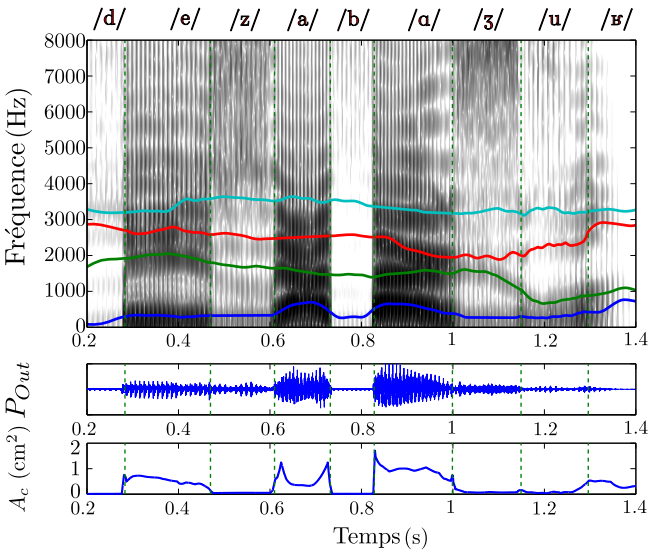

FIGURE 6 – Spectrogramme large-bande, signal de pression acoustique, et aire minimale de la constricition supraglottique, issus de la simulation de la phrase "Des abat-jours" (dezabɑʒuʁ). La segmentation phonétique est indiquée par des lignes traitillées verticales. Les formants estimés à partir du signal du locuteur sont indiqués par des lignes pleines sur le spectrogramme.

Les indices acoustiques tels que la trajectoire formantique, la segmentation phonétique, et l'enveloppe temporelle, du signal de parole synthétisé à partir des données extraites des images IRM, et représenté en figure 6, correspondent bien à ceux de la phrase "Des abat-jours" (dezabɑʒuʁ) prononcée originellement par le locuteur. On remarque également que l'évolution de l'aire minimale de la constricition supraglottique au cours du temps respecte la segmentation phonétique : l'aire est très petite pour les

fricatives /z/ et /ʒ/, et pour la voyelle fermée /u/. Elle est nulle pour les occlusives /d/ et /b/.

6 Conclusion

La méthode présentée pour acquérir des images en 2D du conduit vocal permet d'extraire l'évolution temporelles des contours des articulateurs avec une bonne résolution temporelle. Dans le contexte de la parole naturelle, fixer la cadence d'acquisition autour de 30 images par secondes est un bon compromis entre la qualité d'image et la cadence d'acquisition. Les mouvements des articulateurs, tels que les lèvres et la pointe de la langue peuvent être finement analysés à l'aide d'un algorithme de suivi de contour semi-automatique. Un exemple est présenté dans cet article, montrant que le suivi est en parfait accord avec la segmentation phonétique issue du signal audio. La pertinence des formes géométriques du conduit vocal ainsi extraites est vérifiée par la bonne qualité de la simulation de la propagation acoustique à l'intérieur du conduit vocal articulé selon les contours extraits des films IRM. Ainsi, à l'aide de la connaissance des mouvements articulatoires des locuteurs, et de l'accès à des grandeurs physiques difficilement mesurables (pression intra-orale, débit glottique, mouvement des plis vocaux…) à l'aide d'un synthétiseur articulatoire adapté, il est possible d'étudier en profondeur l'impact de ses mouvements sur les mécanismes de production de la parole, tant du point de vue de la source, que des indices acoustiques produits. Une grande base de données est actuellement alimentée en suivant le protocole expérimentale présenté [1]. Cette base de données est conçue pour, *in fine*, créer à partir de méthodes statistiques un modèle articulatoire représentant la géométrie du conduit vocal de manière réaliste.

Remerciements

Les auteurs remercient FEDER et la région Lorraine pour le soutien financier.

Références

[1] Shrikanth Narayanan, Krishna Nayak, Sungbok Lee, Abhinav Sethy, and Dani Byrd, "An approach to real-time magnetic resonance imaging for speech production," *Journal of the acoustical society of America*, vol. 115, no. 4, pp. 1771–1776, 2004.

[2] Maojing Fu, Bo Zhao, Christopher Carignan, Ryan K Shosted, Jamie L Perry, David P Kuehn, Zhi-Pei Liang, and Bradley P Sutton, "High-resolution dynamic speech imaging with joint low-rank and sparsity constraints," *Magnetic Resonance in Medicine*, vol. 73, no. 5, pp. 1820–1832, 2015.

[3] Sajan Goud Lingala, Yinghua Zhu, Yoon-Chul Kim, Asterios Toutios, Shrikanth Narayanan, and Krishna S Nayak, "A fast and flexible mri system for the study of dynamic vocal tract shaping," *Magnetic resonance in medicine*, 2016.

1. Des fichiers vidéos sont accessibles à l'adresse suivante :`www.loria.fr/~belie/pages/csmri.html`

[4] David L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, pp. 1289–1306, 2006.

[5] Dante C Youla, "Generalized image restoration by the method of alternating orthogonal projections," *Circuits and systems, IEEE transactions on*, vol. 25, no. 9, pp. 694–702, 1978.

[6] Klaas P Pruessmann, Markus Weiger, Markus B Scheidegger, Peter Boesiger, et al., "Sense : sensitivity encoding for fast mri," *Magnetic resonance in medicine*, vol. 42, no. 5, pp. 952–962, 1999.

[7] Mark A Griswold, Peter M Jakob, Robin M Heidemann, Mathias Nittka, Vladimir Jellus, Jianmin Wang, Berthold Kiefer, and Axel Haase, "Generalized autocalibrating partially parallel acquisitions (grappa)," *Magnetic resonance in medicine*, vol. 47, no. 6, pp. 1202–1210, 2002.

[8] D Liang, KF King, B Liu, and L Ying, "Accelerating sense using distributed compressed sensing," in *Proc Intl Soc Mag Reson Med*, 2009, vol. 17, p. 377.

[9] Douglas C Noll, Dwight G Nishimura, and Albert Macovski, "Homodyne detection in magnetic resonance imaging," *Medical Imaging, IEEE Transactions on*, vol. 10, no. 2, pp. 154–163, 1991.

[10] F. Krahmer and R. Ward, "Stable and robust sampling strategies for compressive imaging," *Image Processing, IEEE Transactions on*, vol. 23, no. 2, pp. 612–622, Feb 2014.

[11] E. van den Berg and M. P. Friedlander, "SPGL1 : A solver for large-scale sparse reconstruction," June 2007, http ://www.cs.ubc.ca/labs/scl/spgl1.

[12] E. van den Berg and M. P. Friedlander, "Probing the pareto frontier for basis pursuit solutions," *SIAM Journal on Scientific Computing*, vol. 31, no. 2, pp. 890–912, 2008.

[13] M. Maggioni, V. Katkovnik, K. Egiazarian, and A. Foi, "A nonlocal transform-domain filter for volumetric data denoising and reconstruction," *IEEE Trans. Image Process.*, vol. 22(1), pp. 119–133, 2013.

[14] Tao Zhang, John M Pauly, Shreyas S Vasanawala, and Michael Lustig, "Coil compression for accelerated imaging with cartesian sampling," *Magnetic Resonance in Medicine*, vol. 69, no. 2, pp. 571–582, 2013.

[15] Pierre-André Vuissoz, Freddy Odille, Yves Laprie, Emmanuel Vincent, and Jacques Felblinger, "Sound synchronization and motion compensated reconstruction for speech cine mri," in *ISMRM 2015 Annual Meeting*, 2015.

[16] J. Fontecave Jallon and F. Berthommier, "A semi-automatic method for extracting vocal-tract movements from x-ray films," *Speech Communication*, vol. 51, no. 2, pp. 97–115, 2009.

[17] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20(4), pp. 1118–1133, 2012.

[18] Benjamin Elie and Yves Laprie, "Extension of the single-matrix formulation of the vocal tract : consideration of bilateral channels and connection of self-oscillating models of vocal folds with glottal chink," Sept. 2015.

[19] Parham Mokhtari, Hironori Takemoto, and Tatsuya Kitamura, "Single-matrix formulation of a time domain acoustic model of the vocal tract with side branches," *Speech Communication*, vol. 50(3), pp. 179 – 190, 2008.

[20] Lucie Bailly, Nathalie Henrich, and Xavier Pelorson, "Vocal fold and ventricular fold vibration in period-doubling phonation : Physiological description and aerodynamic modelinga)," *J. Acoust. Soc. Am.*, vol. 127(5), pp. 3212–3222, 2010.

[21] B. Elie and Y. Laprie, "Audiovisual to area and length functions inversion of human vocal tract," in *Eusipco, Lisbon*, 2014.

[22] Alain Soquet, Véronique Lecuit, Thierry Metens, and Didier Demolin, "Mid-sagittal cut to area function transformations : Direct measurements of mid-sagittal distance and area with MRI," *Speech Communication*, vol. 36(3), pp. 169–180, 2002.